Provenance Challenges for Earth Science

Curt Tilmes

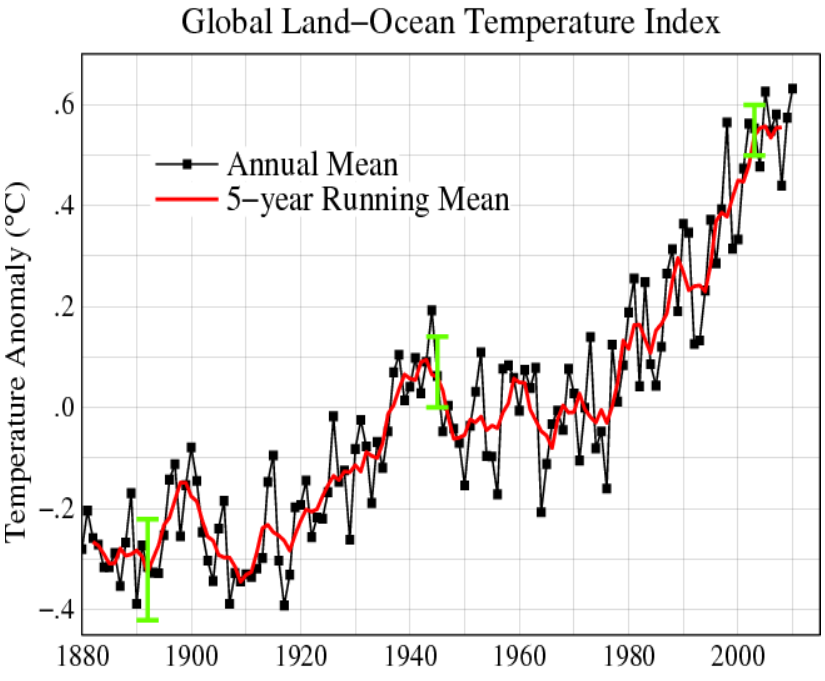
Curt.Tilmes@nasa.gov

OGK 2011
2011-11-04

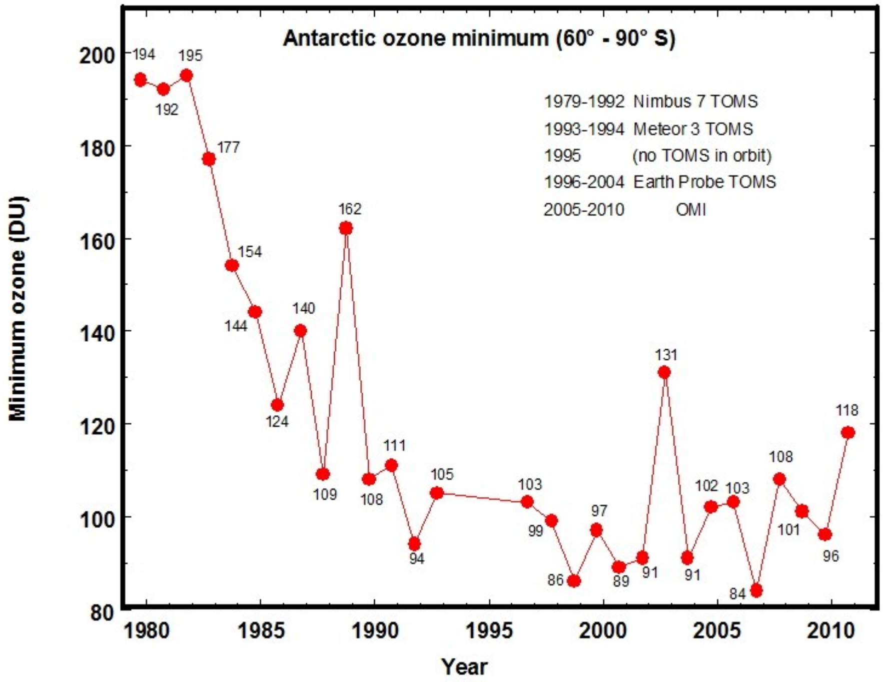
Science

- ❑ “An inherent principle of publication is that others should be able to *replicate* and build upon the authors' published claims. Therefore, a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols available in a publicly accessible database [...] or, where one does not exist, to readers promptly on request.”
  - *(Guide to Publication Policies of the Nature Journals)*
- ❑ Science must be reproducible
  - *(or it isn't science...)*
- ❑ Traditionally, one could read a scientific paper, construct an identical experiment and confirm results
  - *(well, most of the time...)*
- ❑ *Reproducibility* yields *Credibility*

- Some modern scientific research is the result of lengthy computer analysis of a **very large** amount of data, building on the contributions of hundreds (thousands?) of individuals



http://data.giss.nasa.gov/gistemp/graphs/



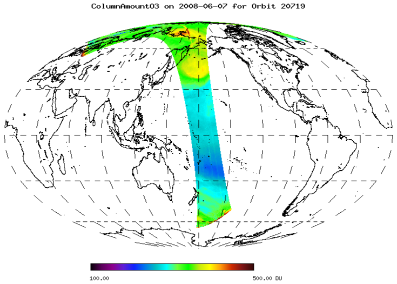
http://jwocky.gsfc.nasa.gov/eptoms/dataqual/ozone_v8.html

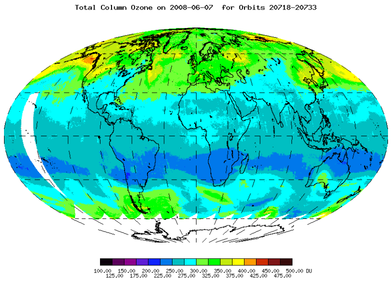
Data Processing and Archiving

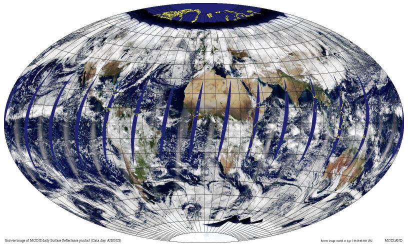
- ❑ Earth Science Data Archive volumes growing steadily
- ❑ Over time, the systems evolve:
  - Spacecraft, sensors, data processing frameworks
  - Science algorithms for transforming and analyzing data
  - Calibration, ancillary lookups
- ❑ Tracking data provenance through processing systems and archives is a very complicated problem
  - Across organizations / agencies this just gets worse
- ❑ Science data is being used in new ways not planned by originators
- ❑ Value Added Services release their own processed data from independent archives
- ❑ Remote web services can be used to transform data

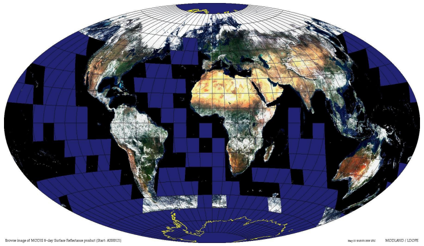
Earth Science Data

- ❑ Data are organized into “granules,” the smallest chunk of data that can be ordered or processed.
- ❑ Examples:
    - OMI Level 2, 1 orbit/granule, 14-15 orbits/day
    - OMI Level 3, 1 day/granule
    - MODIS Level 2 uses 5 minutes of data
    - MODIS Level 3, 1 tile/day/granule,
    - Also, 8 day, 16 day, etc., some are 1 year/granule
- ❑ Each Granule could have multiple arrays and hundreds of metadata fields with information about the data in that granule.
- ❑ A Dataset is all of the granules for a given datatype, distinguished within the dataset by a granule key.
- ❑ In an ongoing mission, new granules are constantly added to a dynamic dataset, both at the end, and occasionally replaced in the middle.






Browse image of MODIS daily Surface Reflectance product (Data day: A2001025)
MODLAND


Browse image of MODIS 8-day Surface Reflectance product (Start: A2000121)
MODLAND / LDOPE

Citations

❑ Current state of practice for citation of Earth Science Datasets is poor to non-existent

- Some have acknowledgements
  - “Thanks to NASA (or NOAA) for data”
  - “Thanks to Fred who gave me some NASA data”
  - “Thanks to MODIS team for MODIS data”
- Some reference specific data inline, with footnotes or in figure captions
  - Used data from Terra MODIS instrument
  - Used Collection 5 Land Surface Reflectance data from Terra MODIS
  - Used Collection 5 Land Surface Reflectance data from Terra MODIS downloaded on 2011-02-08
- A few have started to actually include formal citations in references
  - Even those usually cite the dataset as a whole, not specific granules used in research – that is very difficult for a dynamic dataset.

Citations

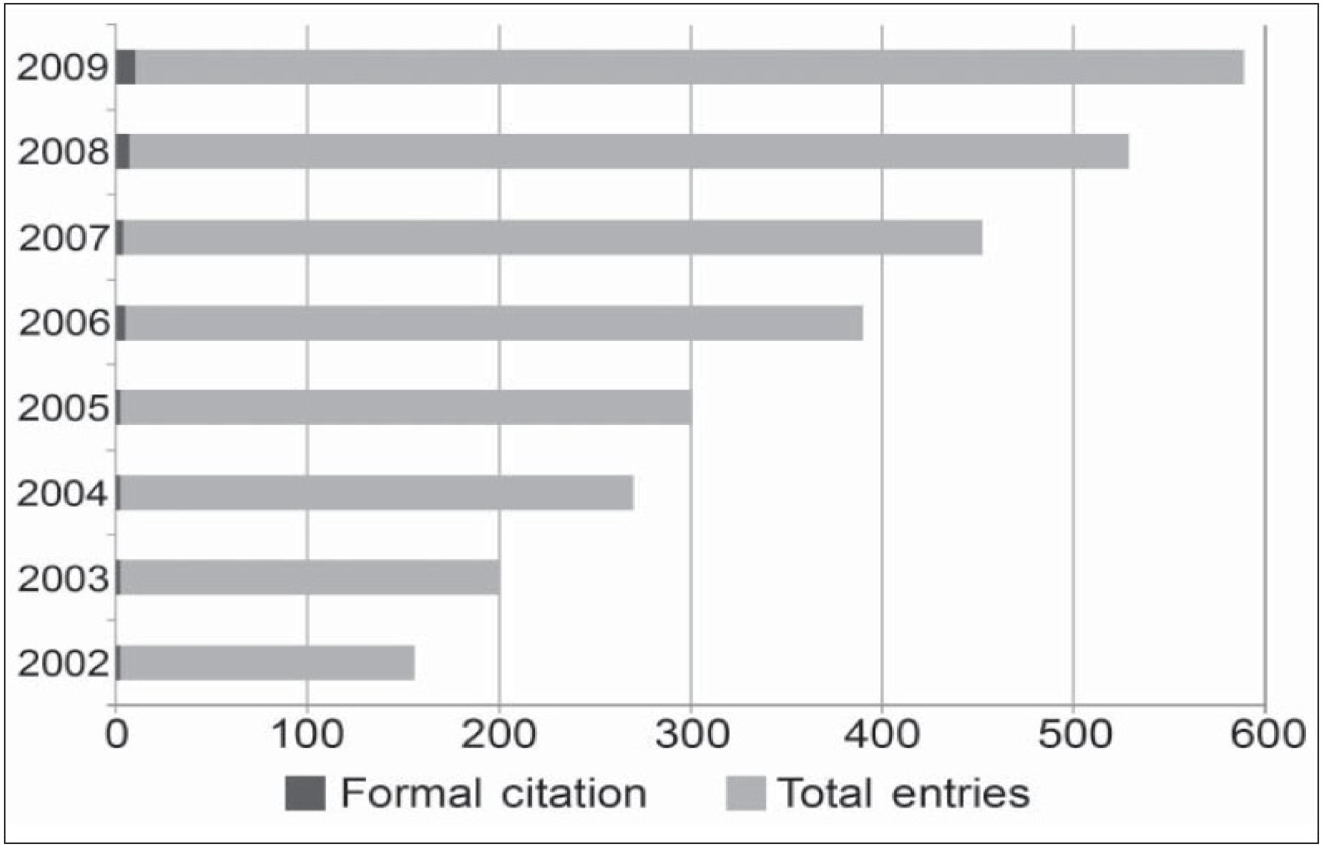


Fig 1. The National Snow and Ice Data Center distributes a variety of different snow cover products derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). The results of a quick analysis of how many scientific papers mention use of "MODIS snow cover data" (according to Google Scholar™) and how often the data sets themselves are formally cited show a huge disparity, illustrating the infrequency of proper data citation in practice. Moreover, the lack of data citation standards introduces the possibility that informal references to data do not point to the data set actually used.

Parsons, et. al. "Data Citation and Peer Review", EOS, Transactions, AGU, 24 Aug. 2010.

When scientific research is published, it should *reference* all data used in that research to a sufficient extent for *others* to *reproduce* that research and confirm the conclusions.

Essential Provenance

- ❑ What aspects of the provenance are "essential" for reproducibility?
- ❑ Some things are definitely "essential"
  - Workflow artifacts – inputs, runtime parameters
- ❑ Some things are definitely "non-essential"
  - Name of processing host, who ran the process, date of processing
  - These are useful for auditing and increase credibility of provenance.
- ❑ Some things aren't so clear
  - Compiler Flags? Library Versions? OS architecture?

Versions

- ❑ Basic configuration management works well for software.

- ❑ Any time the software is changed, we tag a snapshot with a revision number (v. 1.2.3) through our CM tools. – We can go back and check out that version of the software, compare versions, etc.

- ❑ Data versioning is more complicated. The direct predecessors and the software that produced a given granule could have the same version, but due to changes 'up-stream' in the workflow, the data are different.

- ❑ Anytime a new granule is made, it has a distinct identifier, even if it is in the same Dataset with the same Granule Key.

Reprocessing

- Reprocessing – Remake data granules in the best possible way vs.
- Reproduction – Remake a product the “same” way it was made previously.

- We frequently perform large-scale reprocessing with improved algorithms and discard older data – even if they are the basis of published research.

- Operational problems – disk crashes, data lost – reprocess or reproduce?

- Simply delete data that are less used to save disk space, “process-on-demand” when they are wanted

Identical granules

- ❑ For two granules of data to be *Perfectly Identical*, they must not only have identical contents, but also identical identifiers and identical creation provenance. This is only meaningful if you really are talking about the same granule, or two 'copies' of the same granule.

- ❑ Two granules have *Equal Content* if the data contents are the same, even if the identifiers of the granules, or the provenance of the granules are different. It doesn't matter how the content came to be – each such granule can be used in the same analysis and would result in the same results/conclusions.

Equivalent

- ❑ Two granules have *Scientifically Equivalent Content* if the use of those granules in a scientific analysis will lead to the same results or conclusions.
- ❑ This definition allows 'slight' differences in the content – as long as they are close enough not to affect any analysis in a scientifically meaningful way.
- ❑ Proving perfect Scientific Equivalence in the general case is very difficult (impossible?), or at the least, very manual.

Reproducibility

- ❑ *Scientifically Reproducible* refers to a process which is capable of reproducing granules that are *Scientifically Equivalent* to the original granules. *Scientific Reproducibility* is the extent to which a process is *Scientifically Reproducible*.

- ❑ Some processes are chaotic in that very slight differences in processing are compounded possibly producing drastically different results. We can apply sensitivity analyses to assess this characteristic and help determine if the process is suitably reproducible.

- ❑ If a process is unable to reliably reproduce data granules that are *scientifically equivalent*, we would claim that the process is not *reproducible*.

Scientific Equivalence

❑ There are two primary approaches for mechanically approximating this equivalence in a useful way:

- Content Equivalence – Can I show that the contents of two granules are sufficiently equivalent, even if they are not equal?
- Provenance Equivalence – Can I show that two granules were *created* in *essentially* the same way?

Provenance Equivalence

- ❑ We can use a *Provenance Equivalence Identifier* (PEI), created with a digital signature of a canonical serialization of the *essential* provenance of the granule.

- ❑ Each granule sharing a PEI is made in a sufficiently similar manner (they share all *essential provenance* elements) that they are *scientifically equivalent*.

Provenance Equivalence Identifiers

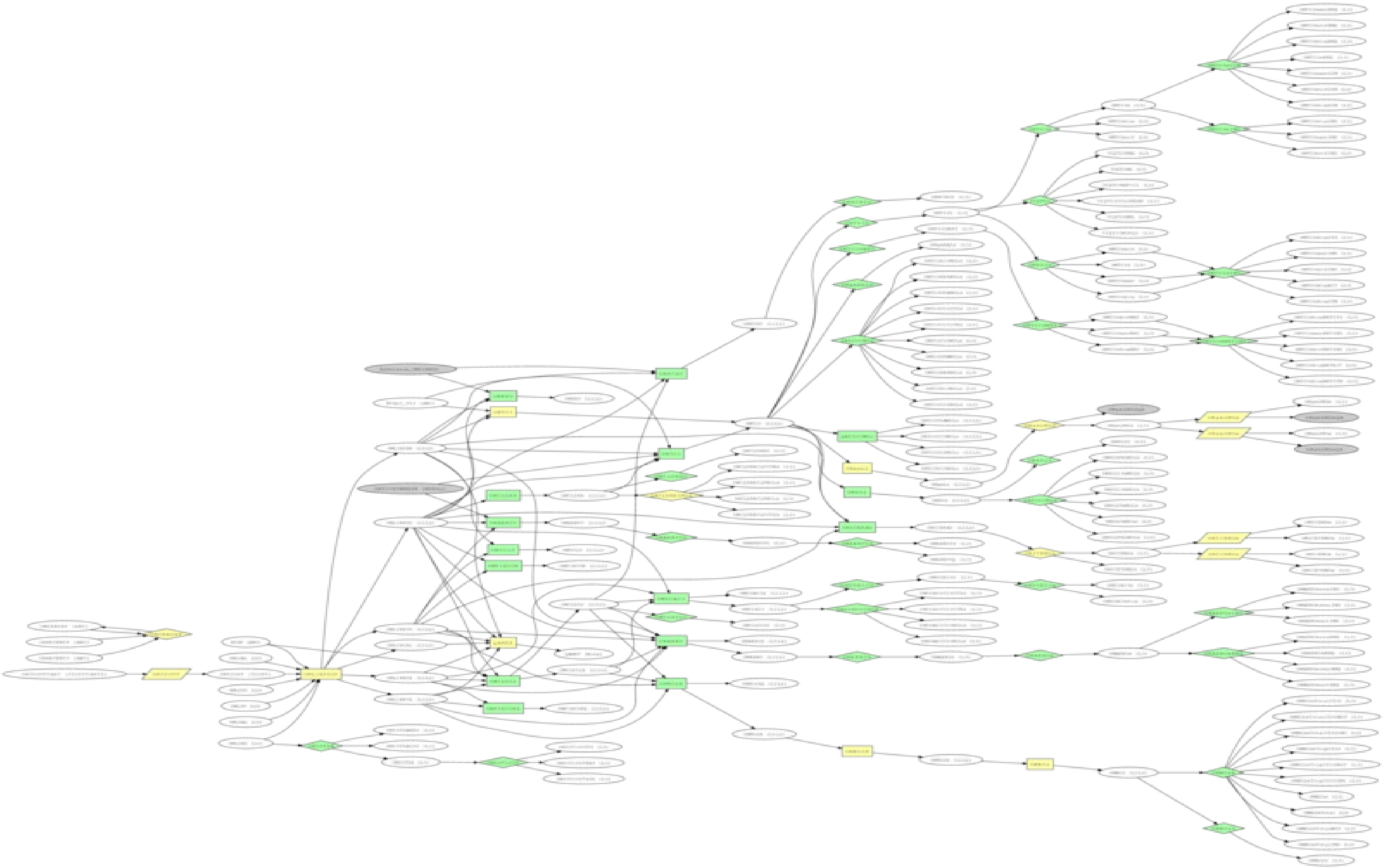
- IF a process is reproducible, we can determine the essential provenance for the process.

- IF we repeat a reproducible process with identical essential provenance, we will get a scientifically equivalent granule.

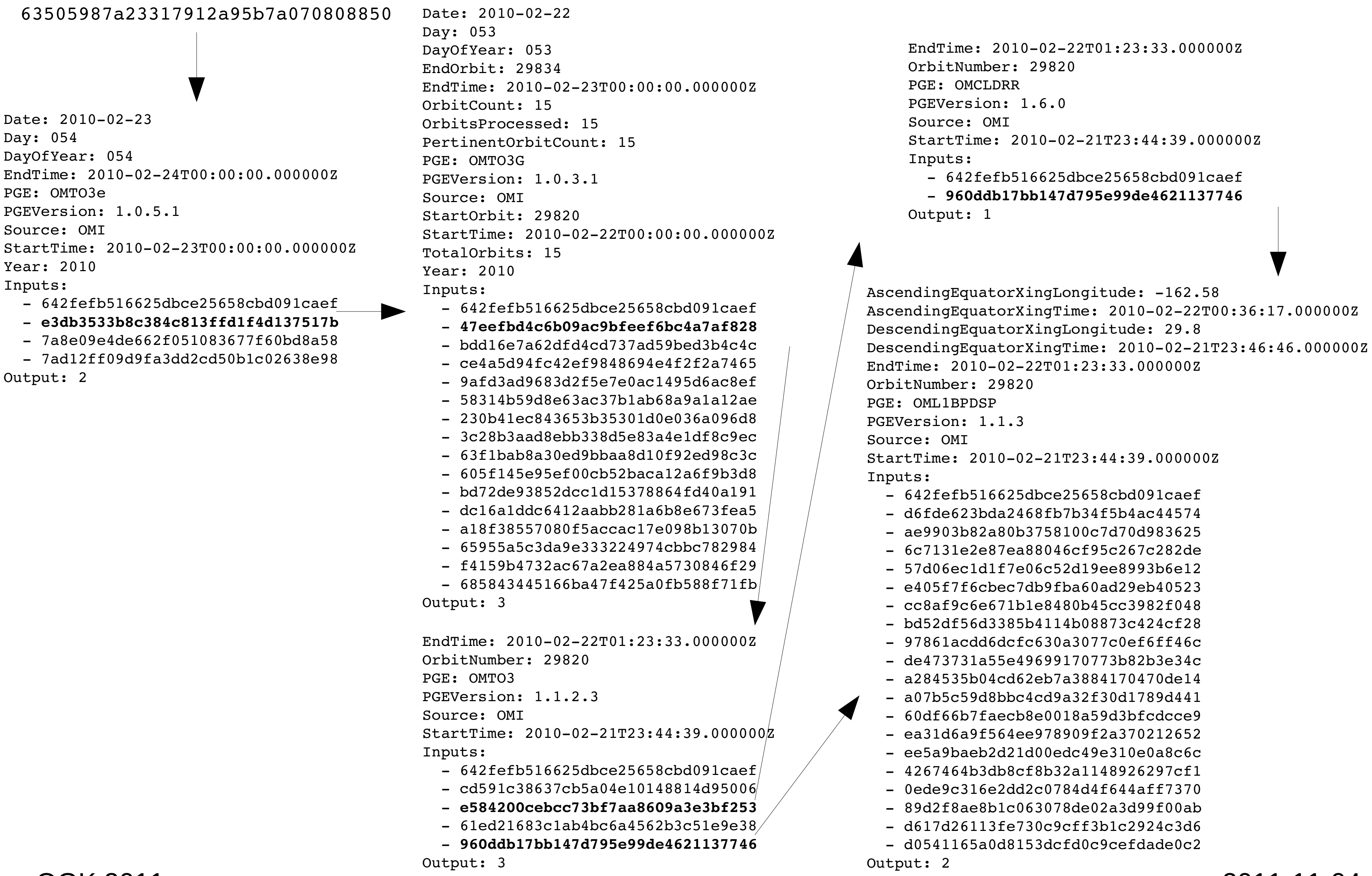
- The PEI can be used as a proxy for the essential provenance graph that led to the creation of that data granule.

- Two granules with the same PEI will be scientifically equivalent to one another, even if their content varies slightly.

OMI Data Flow



PEI Example



```
63505987a23317912a95b7a070808850
```

```
Date: 2010-02-23
Day: 054
DayOfYear: 054
EndTime: 2010-02-24T00:00:00.000000Z
PGE: OMTO3e
PGEVersion: 1.0.5.1
Source: OMI
StartTime: 2010-02-23T00:00:00.000000Z
Year: 2010
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - e3db3533b8c384c813ffd1f4d137517b
  - 7a8e09e4de662f051083677f60bd8a58
  - 7ad12ff09d9fa3dd2cd50b1c02638e98
Output: 2
```

```
Date: 2010-02-22
Day: 053
DayOfYear: 053
EndOrbit: 29834
EndTime: 2010-02-23T00:00:00.000000Z
OrbitCount: 15
OrbitsProcessed: 15
PertinentOrbitCount: 15
PGE: OMTO3G
PGEVersion: 1.0.3.1
Source: OMI
StartOrbit: 29820
StartTime: 2010-02-22T00:00:00.000000Z
TotalOrbits: 15
Year: 2010
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - 47eefbd4c6b09ac9bfeef6bc4a7af828
  - bdd16e7a62dfd4cd737ad59bed3b4c4c
  - ce4a5d94fc42ef9848694e4f2f2a7465
  - 9afd3ad9683d2f5e7e0ac1495d6ac8ef
  - 58314b59d8e63ac37b1ab68a9a1a12ae
  - 230b41ec843653b35301d0e036a096d8
  - 3c28b3aad8ebb338d5e83a4e1df8c9ec
  - 63f1bab8a30ed9bbaa8d10f92ed98c3c
  - 605f145e95ef00cb52baca12a6f9b3d8
  - bd72de93852dcc1d15378864fd40a191
  - dc16a1ddc6412aabb281a6b8e673fea5
  - a18f38557080f5accac17e098b13070b
  - 65955a5c3da9e333224974cbbc782984
  - f4159b4732ac67a2ea884a5730846f29
  - 685843445166ba47f425a0fb588f71fb
Output: 3
```

```
EndTime: 2010-02-22T01:23:33.000000Z
OrbitNumber: 29820
PGE: OMTO3
PGEVersion: 1.1.2.3
Source: OMI
StartTime: 2010-02-21T23:44:39.000000Z
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - cd591c38637cb5a04e10148814d95006
  - e584200cebcc73bf7aa8609a3e3bf253
  - 61ed21683c1ab4bc6a4562b3c51e9e38
  - 960ddb17bb147d795e99de4621137746
Output: 3
```

```
EndTime: 2010-02-22T01:23:33.000000Z
OrbitNumber: 29820
PGE: OMCLDRR
PGEVersion: 1.6.0
Source: OMI
StartTime: 2010-02-21T23:44:39.000000Z
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - 960ddb17bb147d795e99de4621137746
Output: 1
```

```
AscendingEquatorXingLongitude: -162.58
AscendingEquatorXingTime: 2010-02-22T00:36:17.000000Z
DescendingEquatorXingLongitude: 29.8
DescendingEquatorXingTime: 2010-02-21T23:46:46.000000Z
EndTime: 2010-02-22T01:23:33.000000Z
OrbitNumber: 29820
PGE: OML1BPDSP
PGEVersion: 1.1.3
Source: OMI
StartTime: 2010-02-21T23:44:39.000000Z
Inputs:
  - 642fefb516625dbce25658cbd091caef
  - d6fde623bda2468fb7b34f5b4ac44574
  - ae9903b82a80b3758100c7d70d983625
  - 6c7131e2e87ea88046cf95c267c282de
  - 57d06ec1d1f7e06c52d19ee8993b6e12
  - e405f7f6cbec7db9fba60ad29eb40523
  - cc8af9c6e671b1e8480b45cc3982f048
  - bd52df56d3385b4114b08873c424cf28
  - 97861acdd6dcfc630a3077c0ef6ff46c
  - de473731a55e49699170773b82b3e34c
  - a284535b04cd62eb7a3884170470de14
  - a07b5c59d8bbc4cd9a32f30d1789d441
  - 60df66b7faecb8e0018a59d3bfcdcce9
  - ea31d6a9f564ee978909f2a370212652
  - ee5a9baeb2d21d00edc49e310e0a8c6c
  - 4267464b3db8cf8b32a1148926297cf1
  - 0ede9c316e2dd2c0784d4f644aff7370
  - 89d2f8ae8b1c063078de02a3d99f00ab
  - d617d26113fe730c9cff3b1c2924c3d6
  - d0541165a0d8153dcfd0c9cefdade0c2
Output: 2
```

PEI

- ❑ We can follow the provenance equivalence through multiple layers of production.

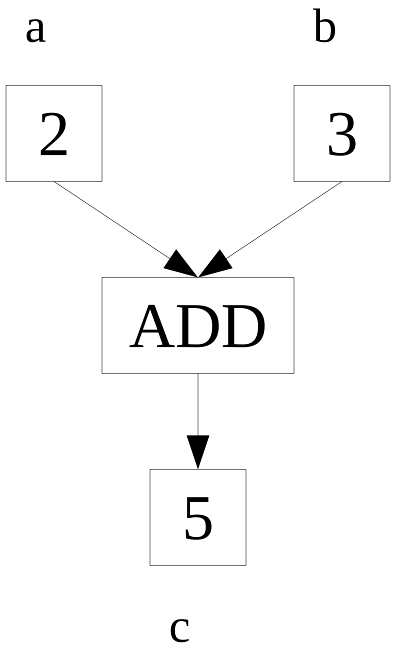
- ❑ Indexing the database on the PEI allows the system to locate equivalent granules.

- ❑ When portions of the data are removed, we can use the metadata and provenance database to determine the “essential provenance” using equivalence of predecessor files rather than requiring the exact files.

- ❑ The system can use “process on demand” to remake previous data, and assert its equivalence to the original.

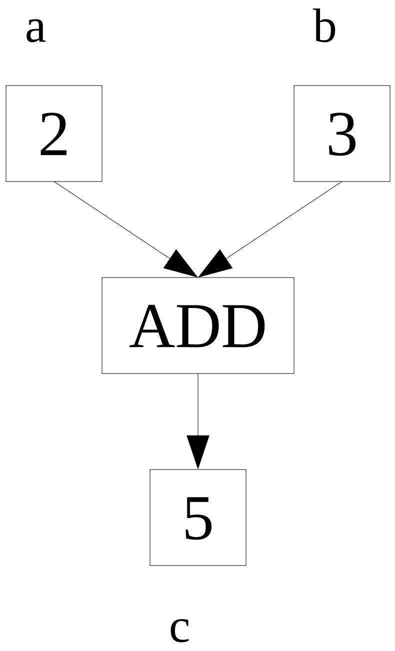
Thank You!

Equivalence of Scientific Data



Granule “c” was created by applying process ADD to input granules “a” and “b”

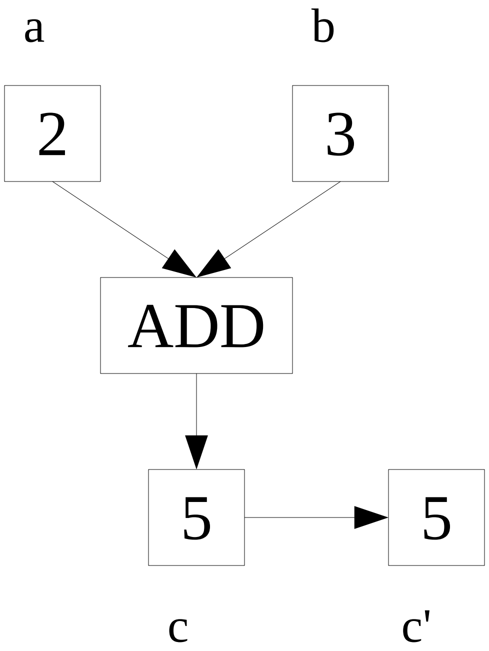
Equivalence of Scientific Data



Granule “c” was created by applying process ADD to input granules “a” and “b”

Joe performed this operation on Feb 2, 2011

Equivalence of Scientific Data



Granule “c” was created by applying process ADD to input granules “a” and “b”

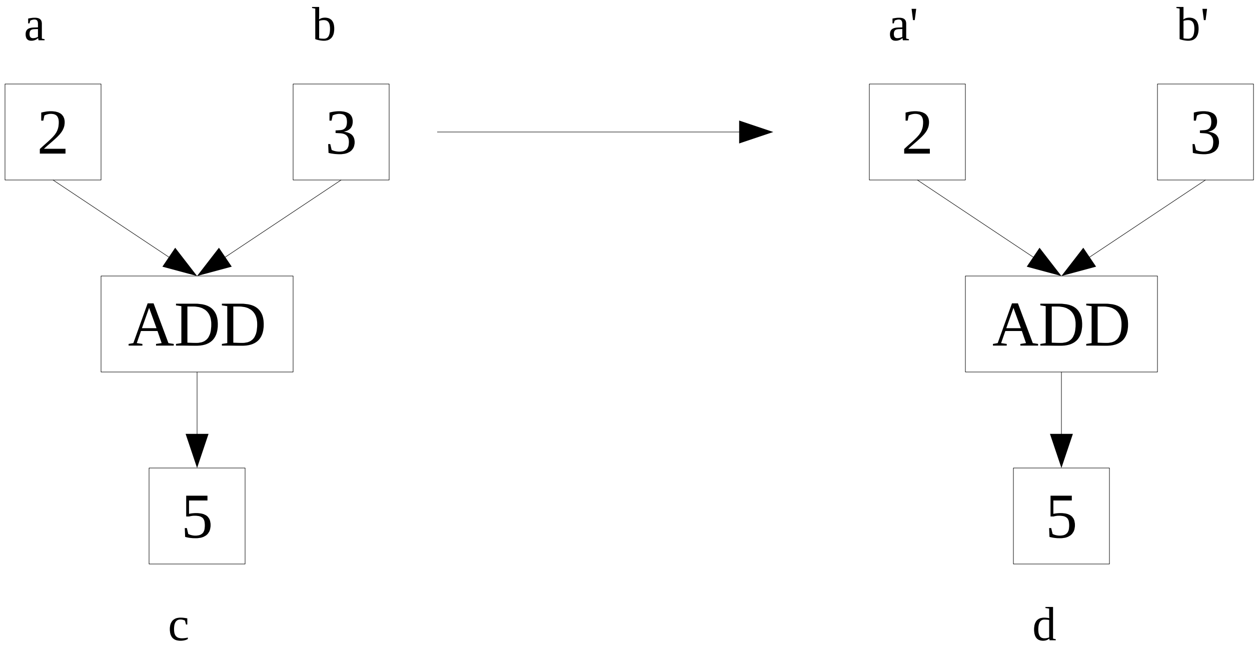
Joe performed this operation on Feb 2, 2011

Fred downloaded granule c from Joe's archive on Feb 5, 2011

c and c' are 'identical' granules.

“creation” provenance vs.
“acquisition provenance”

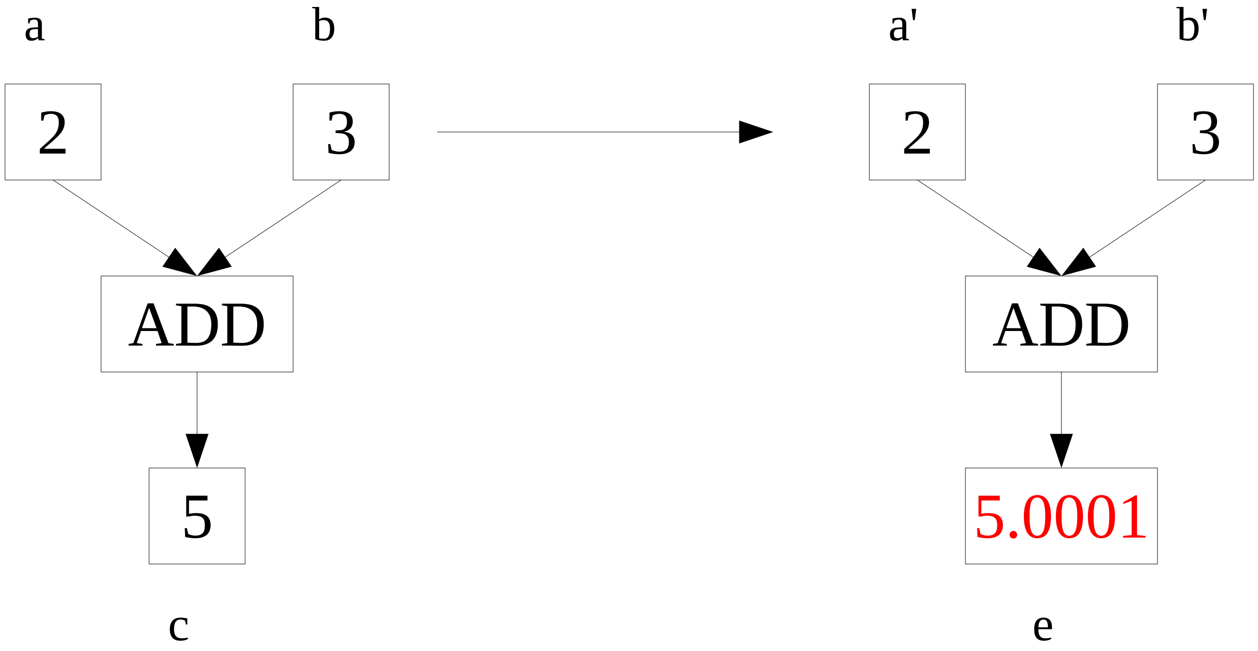
Equivalence of Scientific Data



Sue downloaded a and b, and re-ran process ADD on them producing granule d. She ran this on March 1, 2011.

d has equal content to c

Equivalence of Scientific Data

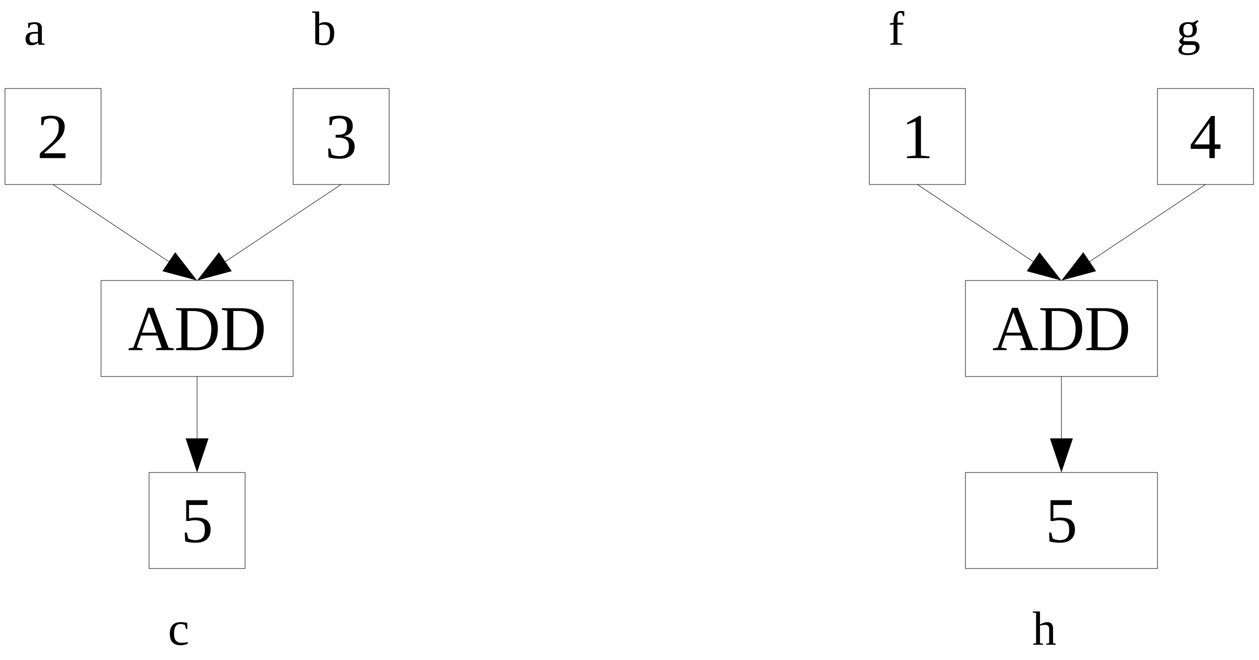


Sue downloaded a and b, and re-ran process ADD on them producing granule e

Her environment has slight differences, so the content is slightly off...

e does not have equal content to c. It *may* be 'equivalent'.

Equivalence of Scientific Data



Sue ran process ADD on input granules f and g producing granule h.

h has equal content to c, but different creation provenance.

(Rarely happens in real life...)

Provenance and Context Artifacts

- ❑ All of the “artifacts” involved or related to the scientific result:
  - Data
  - Algorithms, Processes, Configuration Tables, Runtime Parameters (“Workflow Provenance”)
  - Documentation (ATBDs, Design Docs, Commented Source)
  - Sensors/Instruments/Instrument platforms
  - People/Organizations (reputation)
  - Published scientific papers (add to credibility and understanding)
  - Computer systems, Hardware, OS, Libraries, Software
  - Abstract things like “a data transformation event,” “Software Build Event” or “a validation experiment”
  - An ephemeral execution of a web service
  - Versions from all of the above: Rigorous Configuration Management.
  - Specific relationships between all the artifacts.
- ❑ Things that increase *understanding* and enable *reproducibility*.

Leaf Granules

❑ Some granules come from 'outside' our processing system's scope. If they already have a PEI assigned to them --- great --- if not, we need to 'prime the pump'.

❑ Calculate a digital signature / hash of the content of the granule, and use that as the PEI.

❑ Independent systems that get the same granule will produce the same PEI for that granule.

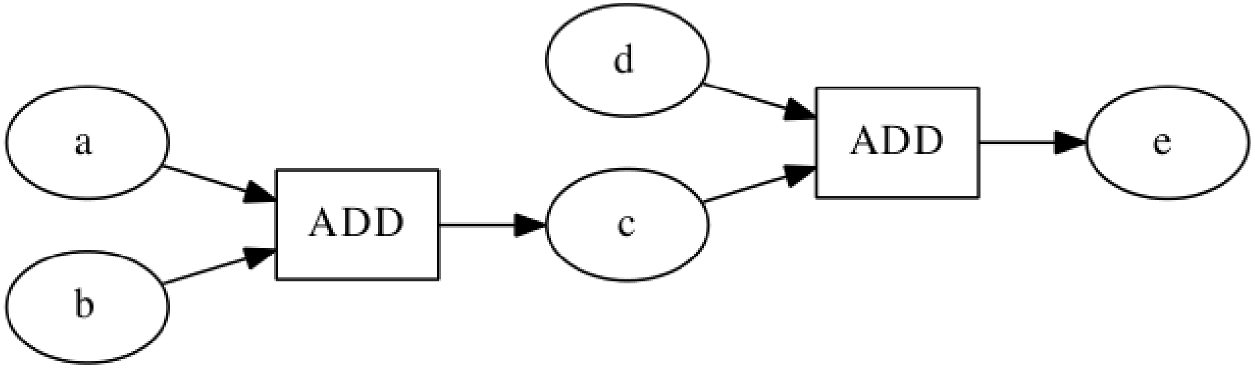
PEI Calculation

- ❑ The PEI for each subsequent data granule is a hash of a canonical serialization of the essential provenance for that granule.

- ❑ For our demonstration implementation, and the examples here, we simplify to three things:
  - Runtime Parameters – these can change the manner of execution of the APP, environment variables, command line arguments, APP identifier, APP version
  - Input Granules – the PEIs of all other input files to the process. The order must be the same.
  - Output Granule Distinguisher – If there are more than one output file, we use a serial number to guarantee a distinct PEI.

PEI Example

- ❑ Simple workflow adding some numbers.



- ❑ a,b,d are leaf granules:

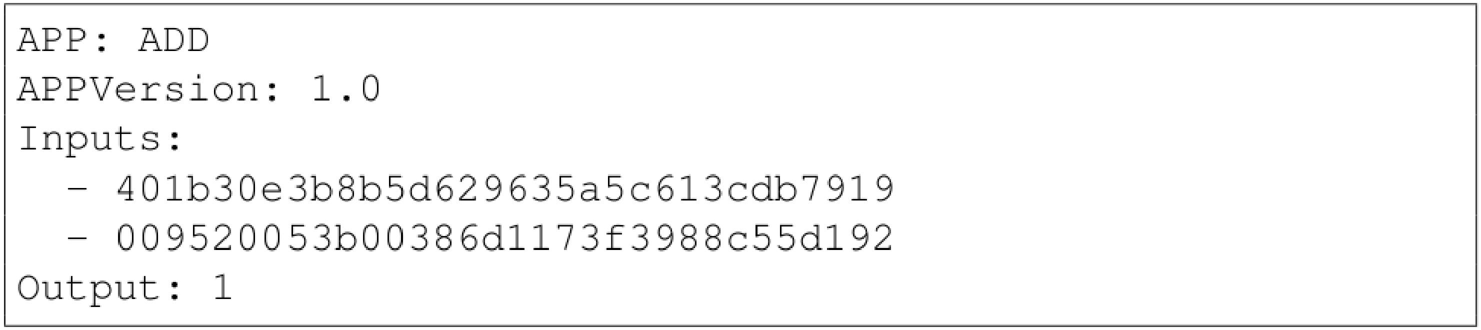
$$\mathbf{PEI}(a) = \texttt{401b30e3b8b5d629635a5c613cdb7919}$$

$$\mathbf{PEI}(b) = \texttt{009520053b00386d1173f3988c55d192}$$

$$\mathbf{PEI}(d) = \texttt{e29311f6f1bf1af907f9ef9f44b8328b}$$

PEI Example (2)

- ❑ Construct a Provenance Equivalence File (PEF) to calculate the PEI of c:

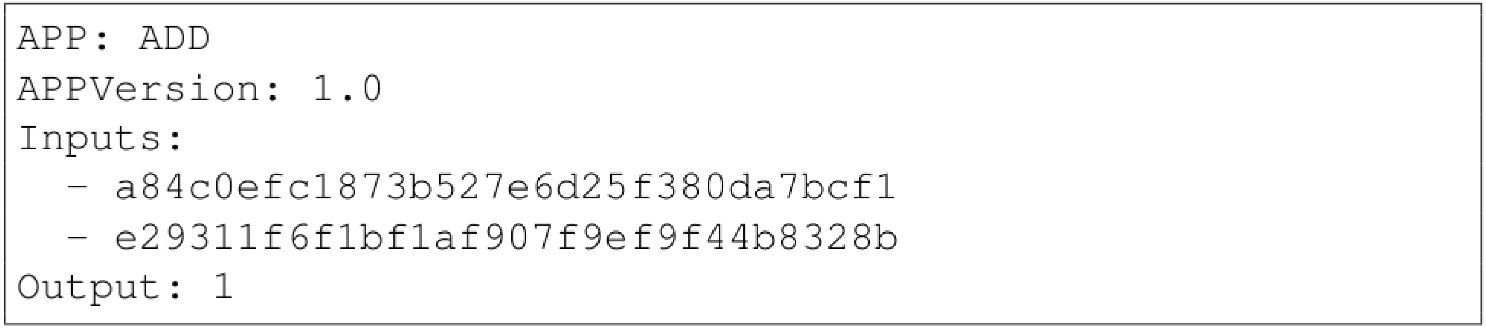


```
APP: ADD
APPVersion: 1.0
Inputs:
  - 401b30e3b8b5d629635a5c613cdb7919
  - 009520053b00386d1173f3988c55d192
Output: 1
```

$$\mathrm{PEI}(c) = \texttt{a84c0efc1873b527e6d25f380da7bcf1}$$

PEI Example (3)

- Construct a PEF and calculate the PEI of e:



```
APP: ADD
APPVersion: 1.0
Inputs:
  - a84c0efc1873b527e6d25f380da7bcf1
  - e29311f6f1bf1af907f9ef9f44b8328b
Output: 1
```

$$\mathrm{PEI}(e) = \texttt{cbedcb426502400ecf4f40a7dd7de89f}$$